Language Identification and Morphosyntactic Tagging: The Second VarDial Evaluation Campaign

Marcos Zampieri[1], Shervin Malmasi[2], Preslav Nakov[3], Ahmed Ali[3], Suwon Shon[4] James Glass[4], Yves Scherrer[5], Tanja Samardžić[6], Nikola Ljubešić[7,8], Jörg Tiedemann[5] Chris van der Lee[9], Stefan Grondelaers[10], Nelleke Oostdijk[10], Dirk Speelman[11] Antal van den Bosch[10,12], Ritesh Kumar[13], Bornini Lahiri[14], Mayank Jain[15]

[1]University of Wolverhampton, [2]Harvard Medical School
[3]Qatar Computing Research Institute, HBKU, [4]Massachusetts Institute of Technology (MIT)
[5]University of Helsinki, [6]University of Zurich, [7]Jožef Stefan Institute, [8]University of Zagreb
[9]Tilburg University, [10]Radboud University, [11]University of Leuven, [12]Meertens Institute
[13]Bhim Rao Ambedkar University, [14]Jadavpur University, [15]Jawaharlal Nehru University

Abstract

We present the results and the findings of the Second VarDial Evaluation Campaign on Natural Language Processing (NLP) for Similar Languages, Varieties and Dialects. The campaign was organized as part of the fifth edition of the VarDial workshop, collocated with COLING'2018. This year, the campaign included five shared tasks, including two task re-runs – Arabic Dialect Identification (ADI) and German Dialect Identification (GDI) –, and three new tasks – Morphosyntactic Tagging of Tweets (MTT), Discriminating between Dutch and Flemish in Subtitles (DFS), and Indo-Aryan Language Identification (ILI). A total of 24 teams submitted runs across the five shared tasks, and contributed 22 system description papers, which were included in the VarDial workshop proceedings and are referred to in this report.

1 Introduction

The interest in applying Natural Language Processing (NLP) methods to similar languages, varieties, and dialects has been growing in recent years. This is evidenced by the growing number of publications and the organization of well-attended workshops co-located with the major NLP conferences such as LT4CloseLang at EMNLP'2014 and the now well-established VarDial workshop series, which is currently in its fifth edition and has been co-located with conferences such as COLING and EACL.

Since its first edition, shared tasks have been organized as part of VarDial. The Discriminating Between Similar Languages (DSL) shared task (Zampieri et al., 2014) was run continuously from 2014 to 2018. In 2016, the DSL task was split into two sub-tasks: a second iteration of the DSL task and the first iteration of the Arabic Dialect Identification (ADI) shared task (Malmasi et al., 2016). In the following year, the organizers decided to broaden the scope of the workshop and to organize an evaluation campaign with four shared tasks (Zampieri et al., 2017): along with iterations of the ADI and the DSL shared tasks, new tasks were started such as the first German Dialect Identification (GDI) and the shared task on Cross-lingual Dependency Parsing (CLP). This year, we continue with a similar setup, covering five shared tasks as part of the Second VarDial Evaluation Campaign.

The remainder of this paper is organized as follows: Section 2 describes this year's shared tasks, Section 3 presents the teams who participated in each task including references to their system descriptions, Section 4 briefly summarizes the related work on the topics of the campaign and on the previous iterations of the ADI and GDI shared tasks. Sections 5, 6, 7, 8, and 9, present the data, the task setup, and the results for each of the shared tasks. Finally, Section 10 concludes this report and points to possible directions for future work.

Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects, pages 1–17
Santa Fe, New Mexico, USA, August 20, 2018.

2 Shared Tasks at VarDial 2018

The VarDial Evaluation Campaign 2018 featured five shared tasks including two task re-runs and three new shared tasks. The two task re-runs were the following:

Third Arabic Dialect Identification (ADI): This year's third edition of the ADI task addressed the multi-dialectal challenge in spoken Arabic in the broadcast domain. Previously, we shared acoustic features and lexical word sequences extracted from large-vocabulary speech recognition (LVCSR). This year, we added phonetic features, aiming at enabling the use of both prosodic and phonetic features, which are helpful for distinguishing between different dialects. We have seen many researchers combine acoustic with linguistic features in previous years (Malmasi et al., 2016; Zampieri et al., 2017), and thus we thought it would be interesting to explore the contribution of phonetic features in overall dialect identification systems.

Second German Dialect Identification (GDI): Following a successful first edition of the (Swiss) German Dialect Identification task in 2017, a second iteration of the GDI task has been organized. We provided updated data on the same Swiss German dialect areas as last year (Basel, Bern, Lucerne, Zurich), and added a fifth "surprise dialect", for which no training data was made available. The participants could take part in two sub-tracks: one on the traditional four-way classification (without the surprise dialect), and another one on five-way classification (with the surprise dialect).

Along with the two task re-runs, the VarDial evaluation campaign included three new shared tasks:

Morphosyntactic Tagging of Tweets (MTT): This task focused on morphosyntactic annotation (900+ labels) of non-canonical Twitter varieties for three South-Slavic languages: Slovene, Croatian, and Serbian. Task participants obtained large manually annotated and raw canonical datasets, as well as small manually annotated Twitter datasets. The task allowed participants to exploit the varieties on two dimensions: (*i*) a comparison of canonical vs. non-canonical language, and (*ii*) the overall proximity of the three languages.

Discriminating between Dutch and Flemish in Subtitles (DFS): The task focused on determining whether a text is written in the Netherlandic vs. the Flemish variant of the Dutch language. For this task, participants were provided with a dataset consisting of over 50,000 subtitle phrases. Since there is a lack of automatic classification studies on the Netherlandic and the Flemish Dutch varieties, and no Netherlandic/Flemish corpus of this size existed, we believe the task was a scientifically interesting step towards developing and comparing language variety classification models using subtitles, and thereby analyzing the proximity of the language varieties in a new way. The latter is not only of interest for improving computational linguistics applications, but it also adds to insights in variational linguistics in general.

Indo-Aryan Language Identification (ILI): This task focused on identifying five closely-related languages from the Indo-Aryan language family – Hindi (also known as Khari Boli), Braj Bhasha, Awadhi, Bhojpuri and Magahi. These languages are part of a continuum starting from Western Uttar Pradesh (Hindi and Braj Bhasha) to Eastern Uttar Pradesh (Awadhi and Bhojpuri) and the neighbouring Eastern state of Bihar (Bhojpuri and Magahi). For this task, the participants were provided with a dataset of approximately 15,000 sentences in each language, mainly from the literature domain, which were published either on the web or in print. This is the first dataset made available for these languages (except for Hindi). We believe that it will not only be useful for the automatic identification of these languages and for developing NLP applications, but it will also enable insights into the proximity level of these languages, which are often mistakenly considered as varieties of Hindi, especially outside the scholarly linguistic circles.

3 Participating Teams

The VarDial Evaluation Campaign received a very positive response from the NLP community. A total of 54 teams registered to participate in the five shared tasks, which is an absolute record for VarDial. Eventually, 24 teams submitted runs and 22 of them also contributed system description papers. The participants were free to participate in one or more tasks, and the number of submissions varied widely across the tasks, ranging from 6 entries for ADI and MTT to 12 entries for DFS. Table 1 lists the participating teams, the shared tasks they took part in, and a reference to the system description paper.

Team	ADI	DFS	GDI	ILI	MTT	System Description Papers
Arabic_Identification	✓					
benf		✓				
BZU	✓					(Naser and Hanani, 2018)
CLiPS		✓				(Kreutz and Daelemans, 2018)
CEA List DeepLIMA					✓	(Meftah and Semmar, 2018)
CoAStaL					✓	
DFSlangid		✓				
dkosmajac		✓	✓	✓		
GDI_classification			✓			(Ciobanu et al., 2018a)
ILIdentification				✓		(Ciobanu et al., 2018b)
JANES					✓	(Ljubešić, 2018)
JSI					✓	(Ljubešić, 2018)
LaMa		✓	✓	✓		
LTL-UDE					✓	
mmb_lct		✓				(Kroon et al., 2018)
safina	✓	✓	✓	✓		(Ali, 2018a; Ali, 2018b; Ali, 2018c)
STEVENDU2018		✓				(Du and Wang, 2018)
SUKI		✓	✓	✓		(Jauhiainen et al., 2018a; Jauhiainen et al., 2018b; Jauhiainen et al., 2018c)
SYSTRAN	✓					(Michon et al., 2018)
Taurus		✓				(van Halteren and Oostdijk, 2018)
Tübingen-Oslo	✓	✓	✓	✓		(Çöltekin et al., 2018)
Twist Bytes Meta			✓			(Benites et al., 2018)
UH&CU					✓	(Silfverberg and Drobac, 2018)
UnibucKernel	✓					(Butnaru and Ionescu, 2018)
we_are_indian				✓		(Gupta et al., 2018)
XAC		✓	✓	✓		(Barbaresi, 2018)
Total	**6**	**12**	**8**	**8**	**6**	**22**

Table 1: The teams that participated in the VarDial'2018 evaluation campaign.

4 Previous Shared Tasks

Since the first DSL challenge, the shared tasks organized within the scope of the VarDial workshop have enjoyed substantial increase in the number of participants and in the overall interest from the NLP community. This motivated the organizers to turn the shared tasks at VarDial into a more comprehensive evaluation exercise with four shared tasks in 2017. This year, the VarDial workshop featured the second edition of the VarDial evaluation campaign with five shared tasks.

This year's second edition of the VarDial Evaluation Campaign was preceded by the first edition of the campaign in 2017 with four shared tasks (Zampieri et al., 2017). Earlier editions of the VarDial workshop featured the DSL shared task, and the ADI shared tasks, which focused on discriminating between similar languages and language varieties in a multilingual dataset and for Arabic dialects, respectively (Zampieri et al., 2014; Zampieri et al., 2015; Malmasi et al., 2016).

4.1 Previous ADI tasks

The first ADI task was introduced in 2016 (Malmasi et al., 2016). It offered as input only lexical information extracted from Arabic LVCSR. The second iteration of the ADI task (Zampieri et al., 2017) introduced multi-modality for dialect identification, using i-vectors for the acoustic representation in addition to lexical features.

The Arabic Multi-Genre Broadcast MGB-3 challenge (Ali et al., 2017) built on the success of the previous two VarDial ADI tasks and introduced the challenge to the speech community, where more attention was paid to the raw audio data using various acoustic representations as well as unsupervised techniques. It is worth noting that all previous challenges used data from the broadcast news domain, with ten hours per dialect for training and two hours per dialect for development and two hours for testing. In contrast, we had fifty hours for training, ten hours testing, and ten hours for development.

4.2 Previous GDI task

The previous GDI task was part of the first VarDial evaluation campaign (Zampieri et al., 2017). It provided manual transcriptions of recorded interviews from four dialect areas of the German-speaking Switzerland, namely Bern, Basel, Lucerne, and Zurich. The training and the test data was extracted from the ArchiMob corpus (Samardžić et al., 2016). The training data consisted in 3,000–4,000 utterances from 3–5 different speakers per dialect; the test data consisted of about 900 utterances by a single speaker per dialect. A total of ten teams participated in the 2017 GDI task and the two best-performing systems (Bestgen, 2017; Malmasi and Zampieri, 2017b) achieved weighted F1-measure of up to 0.66. Transcribers were shown to affect the performance of the systems, e.g., for the Lucerne dialect, whose test set was transcribed by a different person than the training set, recall figures were only around 0.3.

5 Third Arabic Dialect Identification (ADI)

This year's third edition of the ADI task addressed the multi-dialectal challenge in spoken Arabic in the broadcast domain. Last year, in the second edition of the ADI task (Zampieri et al., 2017), we offered the input represented as (*i*) automatic text transcriptions generated using large-vocabulary speech recognition (LVCSR), and (*ii*) acoustic features. This year, we further added phonetic input, which enabled researchers to use both prosodic and phonetic features, which have been shown to be helpful for distinguishing between different dialects (Najafian et al., 2018). We have seen many researchers combine acoustic and lexical features, and thus it was interesting to explore the potential contribution of phonetic features in an overall dialect identification system.

5.1 Dataset

For training and development, we released the same data as for last year's VarDial evaluation campaign (Zampieri et al., 2017). For testing, we prepared two new datasets: (*i*) an in-domain one as in 2017, and (*ii*) an out-of-domain one from YouTube. The duration of the utterances in the YouTube dataset was uniformly distributed between 5 and 30 seconds. We did not inform the participants that there would be an out-of-domain test dataset; we just merged (*i*) and (*ii*) to make a combined test set, but we then evaluated the two parts separately. Each dataset consisted of five Arabic dialects: Egyptian (EGY), Levantine (LEV), Gulf (GLF), North African (NOR), and Modern Standard Arabic (MSA). Detailed statistics about all ADI datasets are shown in Table 2.

For all datasets, we provided to the participants already extracted acoustic features, ASR output, and phonetic features. For acoustic features, we extracted dialect embeddings using an end-to-end dialect identification system, as studies have shown that embeddings from end-to-end models outperform the conventional i-vectors. We used four convolutional layers and two fully connected layers. The parameters for the DNN structure and the training setup were as described in (Shon et al., 2018). We extracted embeddings from the last fully connected layer which was 600-dimensional.

We generated the ASR output using a multi-dialect LVCSR system trained on 1,200 hours for acoustic modeling and on 110 million words for language modeling. More detail about the system, which is the winning system in the MGB-2 challenge, can be found in (Khurana and Ali, 2016).

For the phoneme features, we used the BUT phoneme recognizer (Matejka et al., 2005), which supports languages such as Czech, Russian, Hungarian and English. Despite the language mismatch, the recognizer made predictions for each phoneme label. This is consistent with a previous study that has shown that the Hungarian phoneme recognizer can be useful for Arabic dialect identification (Shon et al., 2017).

	Training		**Development**		**Testing (Broadcast)**		**Testing (YouTube)**	
Dialect	Ex.	Dur.	Ex.	Dur.	Ex.	Dur.	Ex.	Dur.
EGY	3,093	12.4	298	2.0	302	2.0	1,143	5.5
GLF	2,744	10.0	264	2.0	250	2.1	1,147	5.6
LAV	2,851	10.3	330	2.0	334	2.0	1,131	5.5
MSA	2,183	10.4	281	2.0	262	1.9	944	4.6
NOR	2,954	10.5	351	2.0	344	2.1	980	4.8
Total	13,825	53.6	1,524	10.0	1,492	10.1	5,345	26.0

Table 2: The ADI data: examples (Ex.) in utterances, duration (Dur.), in number of hours.

5.2 Participants and Approaches

In this section, we present a short description of the systems that competed in the ADI shared task:

- **UnibucKernel** system (Butnaru and Ionescu, 2018) combines three kernel matrices: one calculated using just the lexical features, another one computed on embeddings, and a combined kernel computed on the phonetic features. The final matrix is the mean of these three matrices. As a classifier, they used Kernel Ridge Regression. The approach is similar to the systems that ranked second and first in the previous two ADI tasks (Ionescu and Popescu, 2016; Ionescu and Butnaru, 2017).

- **Safina** system (Ali, 2018a) accepts a sequence of 256 characters as input in addition to the acoustic embedding vectors. First, the sequence of characters is one-hot encoded, then it is passed to a GRU layer, which is followed by a convolutional layer with different filter sizes ranging from 2 to 7. The convolutional layer is followed by batch normalizations, max-pooling, and dropout layers, and finally a softmax layer. In contrast, the acoustic embedding vectors go directly to another softmax layer. The final output is the average between these two softmax layers, which represents the probability distribution over the labels.

- **BZU** system (Naser and Hanani, 2018) fuses four models, two feed-forward neural networks and two multiclass support vector machines. All models use embedding of size 600 as features, and the training is conducted on the union of the development and of the training data.

- **SYSTRAN** system (Michon et al., 2018) uses a multi-input convolutional neural network. The system first learns character-based embeddings of sentences and phoneme-based embeddings of phoneme representations by running one-dimension convolutions and max-pooling with various filter sizes. Subsequently, it concatenates the output and the given acoustic embeddings for the sentences. Then, it adds several fully-connected layers, and finally makes a prediction.

- **Tübingen-Oslo** system (Çöltekin et al., 2018) is trained on word and character n-grams using a single SVM classifier, which is fine-tuned using cross-validation. It is similar to the submissions by the same authors to previous VarDial shared tasks (Çöltekin and Rama, 2017; Çöltekin and Rama, 2016). They also tried an approach based on RNN, which worked worse.

- **Arabic_Identification** system is based on an ensemble of SVM classifiers trained on character and word n-grams. The approach is similar to the systems ranked second and first in the previous two ADI tasks (Malmasi and Zampieri, 2017a; Malmasi and Zampieri, 2016).

5.3 Results

Six teams submitted runs for the ADI shared task and the results are shown in Table 3. The best result, an F1 score of 0.589, was achieved by *UnibucKernel*,[1] followed by *safina*, with an F1 score of 0.575. The following three teams are tied for the third place as they are not statistically different.

Rank	Team	F1 (Macro)
1	UnibucKernel	0.589
2	safina	0.576
3	BZU	0.534
3	SYSTRAN	0.529
3	Tübingen-Oslo	0.514
4	Arabic_Identification	0.500

Table 3: ADI results: ranked taking statistical significance into account.

5.4 Summary

We introduced multi-phoneme representation for the dialectal data, thus enriching the multi-modal aspect of the dialectal challenge. For the acoustic data, we introduced new dialectal embedding features from an end-to-end dialect identification system. For the evaluation, we used a new surprise test set collected from YouTube, with the aim to test whether the participating systems are robust with respect to new domains, different from broadcast news. The results show that participants did benefit both from the linguistic and from the acoustic features. In the future, we plan to add more data from different domains.

6 Second German Dialect Identification (GDI)

Following the first edition of the (Swiss) German Dialect Identification task in 2017, we organized a second iteration this year. We provided cleaned and updated data on the same Swiss German dialect areas as last year (Basel, Bern, Lucerne, Zurich), and we also added a fifth "surprise dialect" for which no training data was made available. The participants could take part in the traditional four-way classification (without the surprise dialect) or in the five-way classification (with it). We received eight submissions for the four-way classification task, and one submission for the five-way classification task.

6.1 Dataset

As in 2017, we extracted the training and the test datasets from the ArchiMob corpus of Spoken Swiss German. The last publicly available release of the corpus (Samardžić et al., 2016) contains 34 oral history interviews with informants speaking different Swiss German dialects, with nine additional transcriptions currently available, some of which were included in this year's GDI task.

Each interview was transcribed by one of four transcribers, using the writing system "Schwyzertütschi Dialäktschrift" proposed by Dieth (1986). The transcription is expected to show the phonetic properties of the variety, but in a way that is legible for everybody who is familiar with the standard German orthography. Although its objective is to keep track of the pronunciation, Dieth's transcription method is orthographic and partially adapted to the spelling habits in standard German. Therefore, it does not provide the same precision and explicitness as phonetic transcription methods do. Moreover, the transcription choices are dependent on the dialect, the accentuation of the syllables and – to a substantial degree – also the dialectal background of the transcriber. Following the findings of last year's GDI task, we identified several transcriber-specific idiosyncrasies and unified them wherever possible, so that transcriber effects could be reduced.[2] The transcriptions exclusively used lowercase. Also note that Dieth's system is hardly known to laymen, and thus Swiss German data extracted from social media would look fairly different from our transcripts.

[1] The *UnibucKernel* team had the best system for the ADI task in 2017 as well (Ionescu and Butnaru, 2017).

[2] In 2017, the Lucerne dialect achieved recall values of around 30%, and now we increased it to around 45%.

Dialect	Set	Document IDs	Utterances	Tokens	Transcribers
BE	Train	1142, 1170, 1215	3,889	28,558	P, M
	Dev	1121	1,067	7,404	M
	Test	1203	1,191	12,013	A
BS	Train	1044, 1073, 1075	3,349	27,421	A, P
	Dev	1263	1,572	9,544	A
	Test	1224	1,200	9,802	A
LU	Train	1007, 1261, 1008	3,514	29,441	A, P
	Dev	1195	1,079	8,887	P
	Test	1138, 1235	1,186	11,372	A
ZH	Train	1082, 1087, 1143, 1244, 1270, 1055	3,894	28,820	A, P, M
	Dev	1225	940	8,099	M
	Test	1188, 1083	1,175	9,610	A, M
XY	Test	1212	790	8,938	P

Table 4: ArchiMob interviews used for the GDI task. The surprise dialect is labeled XY.

We provided the updated versions of the GDI 2017 training data for training, the updated GDI 2017 test data for development, and yet unseen data for testing (see Table 4).[3] The training set contains utterances from at least three interviews per dialect, and the development and the test sets each contain utterances from at least one other interview per dialect (see Table 4). For the surprise dialect, we provided a slightly smaller test set. We encouraged the participants to include the development data as additional training data in their final systems.

As surprise dialect, we chose a text from the Valais region. The Valais Swiss German dialect is known to be very distinct from the other Swiss German dialects in terms of pronunciation and lexicon (Scherrer and Stoeckle, 2016). The Valais dialect is geographically closest to Bern and Lucerne; linguistically, it is most closely related to the Bern dialect, although it also shares some linguistic features with the Basel and the Lucerne dialects.

6.2 Participants and Approaches

- The **SUKI** submissions are generated using the HeLI method with language models containing only character 4-grams. The HeLI method is based on a product of relative frequencies with a backoff function between different language models, but the backoff function was not used in the submissions for the GDI task. The winning submission used adaptive language models, which were updated while recursively analyzing the test data.

- The **Twist Bytes Meta** system is trained on multiple features such as words and character n-grams (1-7 words and character bigrams). A linear SVM is trained on each feature set and on top of that, a linear SVM meta classifier using cross-validation was trained to gather the predictions.

- **safina** is based on character-level convolutional neural networks. This model accepts a sequence of 256 characters as input. The sequence of characters are one-hot encoded then go to a recurrent GRU layer (which works as an embedding layer), followed by a convolutional layer with different filter sizes ranging from 2 to 7. The convolutional layer is followed by batch normalization, max-pooling, dropout, and finally a softmax layer. The output of the softmax layer represents the probability distribution over the labels.

- **Tübingen-Oslo** submitted one system based on a linear SVM classifier and another one based on an RNN as previously described in Section 5.

[3] For Lucerne, we exchanged parts of the development and of the training data to reduce the transcriber effects seen last year.

- The **LaMa** system is a blend (weighted vote) of eight classifiers being stochastic gradient descent (hinge and modified Huber), multinomial Naïve Bayes, both counts and tf-idf, FastText, and modified Kneser-Ney smoothing. The classifiers were trained using word n-grams (1-6) and character n-grams (1-8). The hyperparameters were determined with cross-validation and searching on the development set.
- **XAC** system is a refined version of the n-gram-based Bayesline system described in last year's XAC submission to the VarDial shared tasks (Barbaresi, 2017), and previously used as a baseline for the DSL shared task (Tan et al., 2014). The XAC team achieved their best results using a Naïve Bayes classifier.
- The **GDI_classification** system is based on an ensemble of multiple SVM classifiers. The system was trained on various word- and character-level features.
- The **dkosmajac** system is based on a normalized Euclidean distance measure. The distances are calculated between a sample and each class profile. The class profiles are generated by selecting the most frequent features for each class, which results in profiles that are of the same length for all the classes.

6.3 Results

For the standard 4-way classification, we received a total of eight submissions presented in Table 5.

Rank	Team	F1 (Macro)
1	SUKI	0.686
2	Twist Bytes Meta	0.646
2	safina	0.645
2	Tübingen-Oslo	0.640
2	LaMa	0.637
3	XAC	0.634
3	GDI_classification	0.620
4	dkosmajac	0.591
	Twist Bytes Meta (5-way)	0.512

Table 5: GDI results: ranked taking statistical significance into account.

This year, the *SUKI* system is the clear winner, achieving significantly higher results than the other seven teams, with an F1 score of 0.686. Four teams ended the competition tied in the second place: *Twist Bytes Meta*, *safina*, *Tübingen-Oslo* and *LaMa*. This year's results are in the same range as last year's, even though the test data is different. With 5 out of 8 teams in an F1 bracket of 0.012, one could argue that a plateau has been reached for this type of transcription data, but the clear win of the *SUKI* system suggests that further improvements can be achieved.

For the extended 5-way classification, we received one submission, which achieved an F1 score of 0.512. It was able to identify the surprise dialect with 22.8% precision and 11.6% recall, suggesting that identifying unseen dialects is still a hard task. The surprise dialect utterances were most often identified as BE or LU, which are the two dialect areas that are geographically closest.

6.4 Summary

In this second iteration of the GDI task, we provided cleaned and updated data from the same source as in 2017. This allowed us to obtain more stable results across dialects. We also launched a surprise dialect task, whose success was limited; a post-submission survey indicated that (prospective) participants mostly lacked time and resources to adapt their systems to such a semi-supervised scenario. Participants also indicated that acoustic data as well as a larger set of dialects would be welcome additions for future iterations of the GDI task.

7 Morphosyntactic Tagging of Tweets (MTT)

The task on morphosyntactic tagging of tweets focused on annotating each token of utterances in non-canonical Twitter varieties of three South-Slavic languages (Slovene, Croatian, and Serbian) with the correct morphosyntactic label out of more than 900 possible ones. The task participants were provided both with large manually annotated and raw canonical datasets, as well as small manually annotated Twitter datasets. Two dimensions of variety could be exploited in the task: (*i*) the dimension of canonical vs. non-canonical language, and (*ii*) the overall proximity of the three languages.

7.1 Datasets

The provided datasets consisted of three types of data: (*i*) standard manually annotated data (`standard.train`), (*ii*) automatically annotated web data (`web.auto`), and (*iii*) Twitter variety manually annotated data (`twitter.*`). The latter were split into train, dev and test sets, with the test data being withheld for the final evaluation. We give an overview of the different datasets (in number of tokens) in Table 6.

The `twitter.*` datasets come from the Janes-Tag manually annotated dataset of Slovene computer-mediated communication (Erjavec et al., 2017) and the ReLDI-NormTagNER-* manually annotated datasets of Croatian (Ljubešić et al., 2017a) and Serbian (Ljubešić et al., 2017b) tweets. These datasets are all similar in size, with around 40 thousand tokens available for training, 8 thousand for development and 20 thousand for testing.

The `standard.train` datasets mostly cover the general domain. While the Slovene and Croatian datasets are similar in size with around 500 thousand tokens, the Serbian dataset is significantly smaller, with just 87 thousand tokens.

The `web.auto` datasets are large web-based datasets: slWac for Slovene (Erjavec et al., 2015), hrWaC for Croatian, and srWaC for Serbian (Ljubešić and Klubička, 2014). They are automatically annotated with state-of-the-art taggers for standard Slovene (Ljubešić and Erjavec, 2016), Croatian, and Serbian (Ljubešić et al., 2016), respectively.

	twitter.train	**twitter.dev**	**twitter.test**	**standard.train**	**web.auto**
Slovene	37,756	7,056	19,296	586,248	895,875,492
Croatian	45,609	8,886	21,412	506,460	1,397,757,548
Serbian	45,708	9,581	23,327	86,765	554,627,647

Table 6: MTT task: size of the datasets (in number of tokens).

7.2 Participants and Approaches

The following teams handed in their system descriptions:

- The **UH&CU** system uses a bidirectional LSTM system for sequence modeling, representing words via word embeddings and character embeddings encoding the character-level word representation with a separate BiLSTM. They use an intriguing approach to emitting tags: they generate tags as character sequences using an LSTM generator in order to handle unknown tags and complex tags (combinations of several tags for one token as a result of conflating tokens in the non-standard varieties).

- The **JSI** system also applies a bidirectional BiLSTM to model the sequence, representing each word via a combination of word embeddings and character-level word representations obtained via character embeddings from a separate BiLSTM. They train the network first on a concatenation of all manually annotated data, and then they tune it only on non-standard (in-domain) data. They also pretrain the character-level BiLSTM word encoder on automatically generated inflectional lexicons from the available automatically annotated web data.

- The **JANES** system uses a conditional random field for sequence labeling, with carefully engineered context-level, word-level, and character-level features. The authors further enrich the representation of each word with Brown clusters that were calculated on the available web data. They train their system on a combination of standard and non-standard data, in which they overrepresent non-standard data by repeating the non-standard instances. They also heavily borrow data between Croatian and Serbian.

- The **DeepLIMA** system uses a BiGRU modeling technique, representing words as a combination of word-level and character-level embeddings. The latter generate a word representation from character-level embeddings with a separate BiGRU. They exploit both the non-standard (in-domain) and the standard (out-of-domain) training data by training the network first on the standard data, and then on the non-standard data.

7.3 Results

The MTT shared task received seven submissions by six teams, each of the teams submitting results for all the three languages of the shared task.

We used token-level accuracy as an evaluation measure, and we ranked the systems, taking statistical significance into account, based on the McNemar test whether the results of two neighbouring submissions are statistically significantly different at the $p < 0.05$ level.

The results are shown in Table 7. We can see that the three best-performing teams, *UH&CU*, *JSI* and *JANES* share the first position in all languages except for Slovene, where the *JANES* team achieved statistically significantly worse results than the two other teams. All other teams performed below the *HunPos* baseline system, which was trained on a concatenation of all the available manually annotated data per language.

Given that the results for three teams are very close to each other, it is reasonable to assume that these results represent the state of the art in morphosyntactic annotation.

The high ranking of the *JANES* system, which is not neural but CRF-based, has shown that the improvements yielded by using neural networks are rather small. Ljubešić (2018) has also noted that his *JANES* system comes closer to the neural approaches as the level of non-standardness of the test data drops off. Namely, the most similar results between *JANES* and the neural approaches were obtained on Serbian and Croatian, for which 10% and 13% of the tokens, respectively, are non-standard, while significantly lower results were obtained for Slovene, where the percentage of non-standard tokens is about 17%.[4]

	Slovenian		**Croatian**		**Serbian**	
Team	**Acc**	**Rank**	**Acc**	**Rank**	**Acc**	**Rank**
UH&CU	0.884	1	0.887	1	0.900	1
JSI	0.883	1	0.890	1	0.900	1
JANES	0.871	4	0.893	1	0.900	1
CEA List DeepLIMA	0.826	6	0.829	6	0.821	6
LTL-UDE	0.627	7	0.752	7	0.773	7
CoAStaL	0.626	8	0.632	8	0.524	8
HunPos baseline	0.832	5	0.834	5	0.832	5

Table 7: MTT results: ranked taking statistical significance into account.

[4]The author also reports results that he obtained after the system submission deadline, with statistically significant improvements over *JANES* being obtained for all languages, and the improvements strongly correlating with the percentage of non-standard tokens in the test sets.

7.4 Summary

The experimental results have shown that by combining standard (out-of-domain) and non-standard (in-domain) training data, as well as training data from closely related languages by using neural approaches or conditional random fields, the traditional *HunPos* baseline can be beaten by a wide margin, with the error reduction lying somewhere around 45%.

Moreover, the improvements when replacing traditional sequence labeling approaches such as CRFs with corresponding neural ones are quite small. However, conditional random fields need the features to be manually engineered, which requires a good knowledge of the target languages. The results also demonstrate that improvements can be achieved with neural approaches on datasets where the level of non-standardness is highest, showing that, as expected, the more complex modeling approaches start to pay off as the problems get harder.

8 Discriminating between Dutch and Flemish in Subtitles (DFS)

The DSF shared task focused on determining whether a text is written in the Netherlandic or in the Flemish variant of the Dutch language. The participants were provided with professionally produced subtitles written for either a Northern Dutch or a Flemish audience. Since there is a lack of automatic classification studies on Netherlandic and Flemish Dutch varieties, and no Netherlandic/Flemish corpus of this size exists, we believe it is a scientifically interesting step forward to develop and to compare language variety classification using subtitles, and thereby analyze the proximity of the language varieties in a new way. The latter is not only of interest for improving computational linguistics applications, but also for finding insights in variational linguistics in general.

8.1 Dataset

As stated above, the dataset consisted of subtitles from an international media localization company that produces, among others, subtitles for television channels in The Netherlands and Belgium. These subtitles range from documentaries, television shows, and movies. These raw subtitles were originally converted into linguistically annotated text in the original SUBTIEL corpus (van der Lee and van den Bosch, 2017). The dataset used for the current shared task was based on this corpus. A total of 320,500 lines were provided to the participants (300,000 for training, 20,000 for testing, and 500 for development). These lines were randomly taken from the SUBTIEL corpus, while keeping a 50/50 split of Netherlandic and Flemish Dutch lines. Each line consisted of about two to three sentences or parts of a sentence: about the length of a tweet. This resulted in a total of 11,102,274 word tokens for all three sets.

8.2 Participants and Approaches

- **Tübingen-Oslo** team used one system based on a linear SVM classifier and another one based on RNN as previously described in Section 5.
- **Taurus** team used a voting-based system that used character n-grams and n-grams containing syntactical information derived from Frog, Alpino and a custom surfacing procedure.
- **CLiPS** team used an ensemble of two Linear SVMs, one trained on word n-grams and another one trained on part-of-speech n-grams. The prediction for a document was made by the classification method that outputs the highest probability for a label.
- **LaMa** team had a system that is a weighted vote blending 8 classifiers: stochastic gradient descent (hinge and modified huber), multinomial Naïve Bayes, both counts and TF.IDF, FastText, and modified Kneser-Ney smoothing, as previously described in Section 6
- **XAC** team used the "Bayesline" system as described in Section 6. In the DFS shared task, *XAC*'s best result was obtained using a Ridge classifier.
- **safina** team used a one-hot encoded character-level convolutional neural network, based on character-level convolutional neural networks, as previously described in Section 5.

- **STEVENDU2018** team used a Linear SVM trained on word n-grams and a Convolutional Neural Network with pre-trained word embeddings built for Netherlandic and Flemish Dutch each, which were subsequently concatenated.

- **mmb_lct** team used a Naïve Bayes classifier using word unigrams and bigrams.

- **SUKI** team submitted identification results from an identifier using the basic HeLI method with words and character n-grams from 1 to 8. Note that using adaptive language models with the DFS dataset did not improve the results as it did for the GDI and the ILI tasks.

- **DFSlangid** team used n-grams, skip-grams, and clustering-based word representations.

- **dkosmajac** team used normalised Eucledian distance measure using Adaptive Gradient Descent to optimize weights. The features used were character n-grams, as previously described in Section 6.

- **benf** team submitted a system trained on a separate Linear SVM on word and on character n-grams. Then they trained a Linear SVM on the output for the two feature sets.

8.3 Results

The DFS shared task received the highest number of submissions across the five tasks, and it was also the most competitive shared task this year with nine out of twelve teams achieving an F1 score between 0.61 and 0.66. The results are presented in Table 8.

Rank	Team	F1 (Macro)
1	Tübingen-Oslo	0.660
2	Taurus	0.646
3	CLiPS	0.636
3	LaMa	0.633
3	XAC	0.632
3	safina	0.631
4	STEVENDU2018	0.623
4	mmb_lct	0.620
5	SUKI	0.613
6	DFSlangid	0.596
7	dkosmajac	0.567
7	benf	0.558

Table 8: DFS results: ranked taking statistical significance into account.

The best-performing system was the one by *Tübingen-Oslo*, and it achieved an F1 score of 0.66, followed by *Taurus* with 0.646. Four teams: *CLiPS*, *LaMa*, *XAC*, and *safina* ended up tied in the third position. Even though the task proved to be very challenging, all teams achieved scores over 0.5, which is the expected baseline for this task.

8.4 Summary

This year's first DFS Shared Task has shown that discriminating between Dutch and Flemish is a challenging but feasible task: all submissions performed better than a 0.5 baseline. No large differences in performance were found between the groups, but the methods to achieve the best performance were quite different. Supervised methods achieved similar scores with different features, and unsupervised methods were competitive as well. We also received some suggestions from participants to further improve the corpus. It would be interesting to see if and how the performance would differ if the data is updated and cleaned based on this feedback.

9 Indo-Aryan Language Identification (ILI)

Organized for the first time in VarDial 2018, the ILI shared task focused on identifying five closely-related languages from the Indo-Aryan language family: Hindi (also known as Khari Boli), Braj Bhasha, Awadhi, Bhojpuri, and Magahi. These languages form part of a continuum starting from Western Uttar Pradesh (Hindi and Braj Bhasha) to Eastern Uttar Pradesh (Awadhi and Bhojpuri) and the neighbouring Eastern state of Bihar (Bhojpuri and Magahi).

For this task, participants were provided with a dataset of approximately 15,000 sentences in each language, mainly from the literature domain, published on the web or in print. It is the first dataset that is made available for these languages (except for Hindi), and we believe it would be useful not only for automatic identification of languages and for developing NLP applications, but it could also help in gaining insights into the proximity level of these languages, which are hypothesised to form a continuum and are often wrongly considered to be varieties of Hindi, especially outside scholarly linguistic circles.

9.1 Dataset

The data for this task was collected from both hard printed and digital sources. Printed materials were obtained from different institutions that promote these languages. We also gathered data from libraries, as well as from local literary and cultural groups. We collected printed stories, novels and essays in books, magazines, and newspapers. We scanned the printed materials, then we performed OCR, and finally we asked native speakers of the respective languages to correct the OCR output. Since there are no specific OCR models available for these languages, we used the Google OCR for Hindi, part of the Drive API. Since all the languages used the Devanagari script, we expected the OCR to work reasonably well, and overall it did. We further managed to get some blogs in Magahi and Bhojpuri.

There are several corpora already available for Modern Standard Hindi (Kumar, 2012; Kumar, 2014a; Kumar, 2014b; Choudhary and Jha, 2011). However, in order to keep the domain the same as for the other languages, we collected data from blogs that mainly contain stories and novels. Thus, the Modern Standard Hindi data collected for this study is also from the literature domain.[5]

9.2 Participants and Approaches

- The **SUKI** team used HeLI with adaptive language models based on character n-grams from 1 to 6, as previously described in Section 5. The also used an iterative version of the language model adaptation technique, with three additional adaptation epochs.
- **Tübingen-Oslo** team submitted a system using a linear SVM and another one based on an RNN, as previously described in Section 5.
- **XAC** team used the "Bayesline" system, as described in Section 6. In the ILI shared task, XAC's best result was obtained using a Ridge classifier.
- **ILIdentification** team used features such as n-grams, skip-grams, and clustering-based word representations. They tried both single classifiers as well as ensembles and stacked generalization.
- **safina** team used one-hot encoded character-level convolutional neural network, as previously described in Section 6.
- **dkosmajac** used character n-grams with a normalized Eucledian distance measure and Adaptive Gradient Descent to optimize weights, as previously described in Section 6.
- **we_are_indian** team combined an RNN-sequence model with bidirectional LSTMs. They created word embedding for all the languages present in the dataset.
- The **LaMa** team used a Multinomial Naïve Bayes classifier with both word and character n-grams, of size 1-8 and 1-6 respectively, for which the raw counts are the feature values.

[5] For more detail about the dataset, please see (Kumar et al., 2018).

9.3 Results

The ILI shared task received eight submissions, and the results are shown in Table 9.

Rank	Team	F1 (Macro)
1	SUKI	0.958
2	Tübingen-Oslo	0.902
2	XAC	0.898
3	ILIdentification	0.889
4	safina	0.863
5	dkosmajac	0.847
5	we_are_indian	0.836
6	LaMa	0.819

Table 9: ILI results: ranked taking statistical significance into account.

The highest ranked team was *SUKI*, which achieved an F1 score of 0.958, while the *LaMa* team had the lowest F1 score of 0.819. Two teams were tied for the second place: *Tübingen-Oslo* and *XAC*. There was a tie between two teams for the fifth place as well.

9.4 Summary

The first ILI shared task was successful in terms of participation with eight submissions. In terms of performance, all submissions achieved an F1 score of more than 0.81, which is a high score for a 5-class classification set-up, and higher than the results achieved in the other VarDial shared tasks. One of the interesting aspects of the task was the wide variety of approaches used by the participants.

In future work, we plan to increase the dataset of these less-resourced languages: Braj Bhasha, Awadhi, Bhojpuri and Magahi.

10 Conclusion and Future Work

We have presented the results and the findings for the five shared tasks that were organized as part of the VarDial Evaluation Campaign in 2018. Two tasks were re-runs from previous years (ADI and GDI), and there were also three new tasks (DFS, ILI, and MTT).

We included a short description for each participant's systems. For a complete description, we included references to the system description papers, which were presented in the VarDial workshop and published in the workshop proceedings.

The VarDial evaluation campaign was introduced in 2017, following the organization of successful shared tasks that have been co-located with VarDial since 2004. In its second edition, the campaign featured a record number of shared tasks and attracted a record number of participants. Participation in each individual task ranged from six teams competing in the ADI and MTT tasks to 12 teams for the DFS task.

In a potential third edition of the VarDial evaluation campaign, we aim to bring more diversity by organizing competitions on other relevant NLP tasks such as lexical variation or machine translation, to name a few. With the exception of MTT, the shared tasks this year dealt mostly with the problem of discriminating between similar languages, varieties, and dialects. Even though this topic has been attracting a lot of attention from the research community (Jauhiainen et al., 2018d), we believe that there is room for shared tasks on other relevant topics in future iterations of the VarDial evaluation campaign.

Acknowledgements

We would like to thank the participants of the VarDial Evaluation Campaign for their hard work, support, and feedback. We further thank the VarDial workshop program committee members for thoroughly reviewing the shared task system papers as well as this report.

References

Ahmed Ali, Stephan Vogel, and Steve Renals. 2017. Speech recognition challenge in the wild: Arabic MGB-3. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*.

Mohamed Ali. 2018a. Character level convolutional neural network for Arabic dialect identification. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*.

Mohamed Ali. 2018b. Character level convolutional neural network for German dialect identification. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*.

Mohamed Ali. 2018c. Character level convolutional neural network for Indo-Aryan language identification. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*.

Adrien Barbaresi. 2017. Discriminating between similar languages using weighted subword features. In *Proceedings of the VarDial Workshop (VarDial)*.

Adrien Barbaresi. 2018. Computationally efficient discrimination between language varieties with large feature vectors and regularized classifiers. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*.

Fernando Benites, Ralf Grubenmann, Pius von Däniken, Dirk von Grünigen, Jan Deriu, and Mark Cieliebak. 2018. Twist Bytes – German dialect identification with data mining optimization. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*.

Yves Bestgen. 2017. Improving the character ngram model for the DSL task with BM25 weighting and less frequently used feature sets. In *Proceedings of the VarDial Workshop (VarDial)*.

Andrei M. Butnaru and Radu Ionescu. 2018. UnibucKernel Reloaded: First place in Arabic dialect identification for the second year in a row. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*.

Çağrı Çöltekin and Taraka Rama. 2016. Discriminating similar languages with linear SVMs and neural networks. In *Proceedings of the VarDial Workshop (VarDial)*.

Çağrı Çöltekin and Taraka Rama. 2017. Tübingen system in VarDial 2017 shared task: Experiments with language identification and cross-lingual parsing. In *Proceedings of the VarDial Workshop (VarDial)*.

Çağrı Çöltekin, Taraka Rama, and Verena Blaschke. 2018. Tübingen-Oslo team at the VarDial 2018 evaluation campaign: An analysis of n-gram features in language variety identification. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*.

Narayan Choudhary and Girish Nath Jha. 2011. Creating multilingual parallel corpora in Indian languages. In *Proceedings of LTC*.

Alina Maria Ciobanu, Shervin Malmasi, and Liviu P. Dinu. 2018a. German dialect identification using classifier ensembles. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*.

Alina Maria Ciobanu, Marcos Zampieri, Shervin Malmasi, Santanu Pal, and Liviu P. Dinu. 2018b. Discriminating between Indo-Aryan languages using SVM ensembles. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*.

Eugen Dieth. 1986. *Schwyzertütschi Dialäktschrift*. 2 edition.

Steven Du and Yuan Yuan Wang. 2018. STEVENDU2018's system in VarDial 2018: Discriminating between Dutch and Flemish in subtitles. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*.

Tomaž Erjavec, Nikola Ljubešić, and Nataša Logar. 2015. The slWaC corpus of the Slovene web. *Informatica*, 39(1):35–42.

Tomaž Erjavec, Darja Fišer, Jaka Čibej, Špela Arhar Holdt, Nikola Ljubešić, and Katja Zupan. 2017. CMC training corpus Janes-Tag 2.0. Slovenian language resource repository CLARIN.SI.

Divyanshu Gupta, Gourav Dhakad, Jayprakash Gupta, and Anil Kumar Singh. 2018. IIT (BHU) System for Indo-Aryan language identification (ILI) at VarDial 2018. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*.

Radu Tudor Ionescu and Andrei Butnaru. 2017. Learning to identify Arabic and German dialects using multiple kernels. In *Proceedings of the VarDial Workshop (VarDial)*.

Radu Tudor Ionescu and Marius Popescu. 2016. UnibucKernel: An approach for Arabic dialect identification based on multiple string kernels. In *Proceedings of the VarDial Workshop (VarDial)*.

Tommi Jauhiainen, Heidi Jauhiainen, and Krister Lindén. 2018a. HeLI-based experiments in discriminating between Dutch and Flemish subtitles. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*.

Tommi Jauhiainen, Heidi Jauhiainen, and Krister Lindén. 2018b. HeLI-based experiments in Swiss German dialect identification. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*.

Tommi Jauhiainen, Heidi Jauhiainen, and Krister Lindén. 2018c. Iterative language model adaptation for Indo-Aryan language identification. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*.

Tommi Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. 2018d. Automatic language identification in texts: A survey. *arXiv preprint arXiv:1804.08186*.

Sameer Khurana and Ahmed Ali. 2016. QCRI advanced transcription system (QATS) for the Arabic Multi-Dialect Broadcast Media Recognition: MGB-2 Challenge. In *Proceedings of SLT*.

Tim Kreutz and Walter Daelemans. 2018. Exploring classifier combinations for language variety identification. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*.

Martin Kroon, Maria Medvedeva, and Barbara Plank. 2018. When simple n-gram models outperform syntactic approaches: Discriminating between Dutch and Flemish. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*.

Ritesh Kumar, Bornini Lahiri, Deepak Alok, Atul Kr. Ojha, Mayank Jain, Abdul Basit, and Yogesh Dawar. 2018. Automatic identification of closely-related Indian languages: Resources and experiments. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC)*.

Ritesh Kumar. 2012. Challenges in the development of annotated corpora of computer-mediated communication in Indian languages: A case of Hindi. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC)*.

Ritesh Kumar. 2014a. Developing politeness annotated corpus of Hindi blogs. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*.

Ritesh Kumar. 2014b. *Politeness in Online Hindi Texts: Pragmatic and Computational Aspects*. Ph.D. thesis, Jawaharlal Nehru University.

Nikola Ljubešić and Filip Klubička. 2014. {bs,hr,sr}wac - web corpora of Bosnian, Croatian and Serbian. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pages 29–35.

Nikola Ljubešić, Tomaž Erjavec, Maja Miličević, and Tanja Samardžić. 2017a. Croatian Twitter training corpus ReLDI-NormTagNER-hr 2.0. Slovenian language resource repository CLARIN.SI.

Nikola Ljubešić, Tomaž Erjavec, Maja Miličević, and Tanja Samardžić. 2017b. Serbian Twitter training corpus ReLDI-NormTagNER-sr 2.0. Slovenian language resource repository CLARIN.SI.

Nikola Ljubešić. 2018. Comparing CRF and LSTM performance on the task of morphosyntactic tagging of non-standard varieties of South Slavic languages. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*.

Nikola Ljubešić and Tomaž Erjavec. 2016. Corpus vs. lexicon supervision in morphosyntactic tagging: the case of Slovene. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*.

Nikola Ljubešić, Filip Klubička, Željko Agić, and Ivo-Pavao Jazbec. 2016. New inflectional lexicons and training corpora for improved morphosyntactic annotation of Croatian and Serbian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*.

Shervin Malmasi and Marcos Zampieri. 2016. Arabic dialect identification in speech transcripts. In *Proceedings of the VarDial Workshop (VarDial)*.

Shervin Malmasi and Marcos Zampieri. 2017a. Arabic dialect identification using iVectors and ASR transcripts. In *Proceedings of the VarDial Workshop (VarDial)*.

Shervin Malmasi and Marcos Zampieri. 2017b. German dialect identification in interview transcriptions. In *Proceedings of the VarDial Workshop (VarDial)*.

Shervin Malmasi, Marcos Zampieri, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, and Jörg Tiedemann. 2016. Discriminating between similar languages and Arabic dialect identification: A report on the third DSL shared task.

Pavel Matejka, Petr Schwarz, Jan Cernockỳ, and Pavel Chytil. 2005. Phonotactic language identification using high quality phoneme recognition. In *Proc. Interspeech*.

Sara Meftah and Nasredine Semmar. 2018. Using neural transfer learning for morpho-syntactic tagging of South-Slavic languages tweets. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*.

Elise Michon, Minh Quang Pham, Josep Crego, and Jean Senellart. 2018. Neural network architectures for Arabic dialect identification. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*.

Maryam Najafian, Sameer Khurana, Suwon Shon, Ahmed Ali, and James Glass. 2018. Exploiting convolutional neural networks for phonotactic based dialect identification. In *IEEE ICASSP*, pages 5174–5178.

Rabee Naser and Abualsoud Hanani. 2018. Birzeit Arabic dialect identication system for the 2018 VarDial challenge. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*.

Tanja Samardžić, Yves Scherrer, and Elvira Glaser. 2016. ArchiMob – a corpus of spoken Swiss German. In *Proceedings of LREC*.

Yves Scherrer and Philipp Stoeckle. 2016. A quantitative approach to Swiss German – dialectometric analyses and comparisons of linguistic levels. *Dialectologia et Geolinguistica*, 24(1):92–125.

Suwon Shon, Ahmed Ali, and James Glass. 2017. MIT-QCRI Arabic dialect identification system for the 2017 multi-genre broadcast challenge. In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 374–380.

Suwon Shon, Ahmed Ali, and James Glass. 2018. Convolutional neural network and language embeddings for end-to-end dialect recognition. In *Proceedings of the Speaker and Language Recognition Workshop (Odyssey)*.

Miikka Silfverberg and Senka Drobac. 2018. Sub-label dependencies for neural morphological tagging – the joint submission of University of Colorado and University of Helsinki for VarDial 2018. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*.

Liling Tan, Marcos Zampieri, Nikola Ljubešic, and Jörg Tiedemann. 2014. Merging comparable data sources for the discrimination of similar languages: The DSL corpus collection. In *Proceedings of the BUCC Workshop*.

Chris van der Lee and Antal van den Bosch. 2017. Exploring lexical and syntactic features for language variety identification. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*.

Hans van Halteren and Nelleke Oostdijk. 2018. Identification of differences between Dutch language varieties with the VarDial2018 Dutch-Flemish subtitle data. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, and Jörg Tiedemann. 2014. A report on the DSL shared task 2014. In *Proceedings of the VarDial Workshop (VarDial)*.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, Jörg Tiedemann, and Preslav Nakov. 2015. Overview of the DSL shared task 2015. In *Proceedings of the LT4VarDial Workshop (LT4VarDial)*.

Marcos Zampieri, Shervin Malmasi, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, Jörg Tiedemann, Yves Scherrer, and Noëmi Aepli. 2017. Findings of the VarDial evaluation campaign 2017. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*.